COVID-19 VULNERABILITY INDEX

For United States Counties

Capstone Project

Georgia Institute of Technology

Shruthy Nair

Table of Contents

Abstract

COVID-19 is a highly infective virus with a rapid transmission rate. It has led to a pandemic that has impacted millions of people all around the world. In the United States alone, over 3 million people have being directly affected by COVID-19 as they tested positive and millions more have been affected indirectly due to the virus. The purpose of this study is to determine if a COVID-19 Vulnerability Index can be created using GIS, that would enable one to identify high risk counties within the United States. A Vulnerability Index measures how vulnerable a population or region is to a particular illness. Multiple socio-economic, demographic, transportation and health related factors were utilized in the development of the Vulnerability Index. Principal Component Analysis were applied to analyze the distribution and correlation in the factors and create the index values. The COVID-19 case rates, death rates and the COVID-19 Vulnerability Index values were compared using spatial clustering and then their actual results were compared to see if the Vulnerability Index is a good measure for COVID-19 case rates and death rates. Results indicated that the COVID-19 Vulnerability Index is a good measure to identify counties that are at risk of increasing their case rate, but not death rates. Furthermore, ordinary least squares regression and spatial lag model were run to evaluate the effectivity of the COVID-19 Vulnerability Index in identifying counties with increasing risk of COVID-19 cases. The regression models indicated that the Vulnerability Index is a relatively good measure determining high risk counties.

Introduction

There have been nearly 13 million cases of COVID-19 worldwide and over 500,000 deaths. In the United States alone, there are currently over 3 million cases of COVID-19 and over 135,000 deaths. From these statistics alone, it can be concluded that COVID-19 has had a large impact on the world and the United States. COVID-19, which is officially named SARS-CoV-2 or Severe Acute Respiratory Syndrome Coronavirus 2, is an extremely infective illness caused by a virus. The first known instance of the COVID-19 outbreak was identified in December 2019 in the city of Wuhan in Hubei province, China. The swift spread of the virus throughout the world caused the World Health Organization to declare COVID-19 a pandemic on March 11, 2020 and identify it as a Public Health Emergency of International Concern. On March 13, 2020, the United States declared a national emergency due to the COVID-19 outbreak (Centers for Disease Control and Prevention, 2020).

COVID-19 spreads through the air by coughing or sneezing, through close personal contact (including touching and shaking hands) or through touching your nose, mouth or eyes before washing your hands. It is a rapidly spreading disease that puts everyone at risk; older adults and people of varying ages who have other serious underlying medical conditions may be at higher risk for contracting it (Centers for Disease Control and Prevention, 2020). There are several socio-economic, demographic, transportation and health factors that affect the spread of the illness. Researchers and scientists are attempting to understand the factors that affect the spread of COVID-19 as well as the severity of the illness. The purpose of this study is to create and determine if a COVD-19 Vulnerability Index can be utilized to assess the spread of the virus in counties within the United States and identify risk areas. The research question that will be explored in this study is: can a COVID-19 Vulnerability Index be created that would enable one to understand the spread of the virus and predict high risk counties using GIS?

A Vulnerability Index measures the exposure of the population to the illness/hazard (in this case: Coronavirus). It is a process through which one can identify and address factors that cause the spread of Coronavirus (COVID-19) and determine the regions that are at a higher risk. From the initial spread of the COVID-19 in December 2019 to its current state, researchers and scientists have identified and looked at various factors that affect the spread of the virus and the risk to certain population groups. For the purpose of this project, multiple socio-economic, demographic,

transportation and health factors will be considered in order to develop the Vulnerability Index. The results from the index will then be applied to the active cases and death rates to determine if the Vulnerability Index results correlate with the actual results. Based on the level of correlation, the Vulnerability Index can be utilized to predict the risk of COVID-19 to counties. Predicting and identifying counties that are at risk will allow for targeted intervention and control.

Literature Review

In order to identify vulnerable communities (counties) it is important to determine factors that make certain populations or areas more vulnerable to COVID-19 than other groups. Since the prevalence of COVID-19 is so recent, there is relatively less research available with regards to identifying factors that affect the spread and rigorousness of the disease. However, there have been studies that indicated that people who are elderly (above the age of 65) and people who already had pre-existing medical conditions are more vulnerable to COVID-19. Age and pre-existing conditions do not necessarily affect who gets infected by COVID-19, however it does affect how severe there infection will be. COVID-19 has had a more detrimental impact on individuals who were above the age of 65. Based on initial findings in March 2020 (in the United States), 31% of the COVID-19 cases, 45% of the hospitalizations, 53% of the ICU admission, and 80% of the deaths occurred in adults over the age of 65 (Centers for Disease Control and Prevention, 2020). These results were very similar to China, as based on their findings, more than 80% of COVID-19 related deaths were among individuals above the age of 60 (Epidemiology Working Group for NCIP Epidemic Response, 2020).

As mentioned above, individuals who have certain pre-existing health conditions are more susceptible to be severely affected by COVID-19 leading to many potential complications. One has a higher risk of severe illness from COVID-19 if they have chronic kidney disease and are at any stage in the disease. Based on data gathered from various studies, approximately 20% of COVID-19 patients with chronic kidney disease were classified to have severe to extreme complications (Henry & Lippi, 2020). In addition, chronic obstructive pulmonary disease (COPD), coronary artery disease, and heart failure in patients increases the risk of severity of COVID-19 (Centers for Disease Control and Prevention, 2020). Furthermore, a study carried out in the state of New York analyzed pre-existing conditions within 5700 hospitalized COVID-19

patients. Based on their findings, 41.7% and 33.8% of the most common comorbidities in the patients were obesity and diabetes respectively. Ultimately, research and data from several studies indicated that the pre-existing health conditions that were addressed can impact how severe a patient with COVID-19 will be, thus, it is important to include these factors in the creation of the COVID-19 Vulnerability Index.

In addition to age and epidemiological factors, social-economic factors also affect the extent of the COVID-19 disease spread within a region. Within the first 15 weeks of 2020 in Massachusetts, the number of deaths in certain areas had a direct correlation to higher level of poverty and lower income levels in those areas (Ramírez & Lee, 2020). This relationship is also existent in diseases that are similar to COVID-19, such as influenza; where high poverty and low income correlates to increase in vulnerability to the illness (Chandrasekhar, et al., 2017).

Factors pertaining to socio-demographic and commuting aspects also affect the spread and risk of COVID-19. Based on a study conducted in Italy using commuting and demographic census data, in addition to utilizing data pertaining to existing COVID-19 cases, it was determined that 52% of all Italian cases were in municipalities with high population density and where public transportation was a popular method for commuting to and from work, school, etc (Savini, Candeloro, Calistri, & Conte, 2020). The results indicated that human mobility and dense populations in and around cities can majorly affect how fast COVID-19 can spread and the number of people that it can infect. Thus, it is important to consider and include socio-demographic and commuting factors when developing a Vulnerability Index for COVID-19.

Data Sources

The COVID-19 dataset was obtained from the COVID-19 data repository from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University through GitHub. The dataset consists of case and death counts (on a daily basis) for each county in the United States between the dates of January 22, 2020 and June 30, 2020.

Population density data, percent population over the age of 65, federal poverty data, median income and mode of transportation to work data was attained from the United States Census Bureau and the American Community Survey for each county within the United States.

Furthermore, the percent obesity in adults and percent diabetes in adult population data was obtained from the 2020 County Health Rankings based on the prevalence of those conditions in the counties. In addition, the percent population with poor health data was also obtained from the 2020 County Health Rankings. With regards to other health conditions that affect COVID-19 impact such as heart failure, chronic kidney disease, chronic obstructive pulmonary disease and coronary artery disease, it was not possible to attain prevalence data for these conditions on a county level for all ages. Thus, the Centers for Medicare & Medicaid Services data, for prevalence of those conditions in adults over the age of 65 were utilized instead. TIGER shapefiles for each county and state in the United States was obtained from the United States government's data catalog.

Methodology

Before attempting to create a COVID-19 Vulnerability Index, it is important to examine and analyze the COVID-19 dataset for COVID-19 cases and COVID-19 related deaths. Looking at the distribution of the dataset spatially will provide an understanding of any general patterns that are existent in the data. For the purpose of this project, only the 48 mainland states (excluding Alaska and Hawaii) will be considered and utilized for the analysis and results process. In addition, ArcGIS Pro was utilized for analysis, joins, and for the creation of all the maps.

Figure 1 below represents the COVID-19 case rate for each United States County. Based on the results you can notice very high case rates in the southern counties of New York, in Massachusetts, in New Jersey, towards the central regions in the state of Mississippi, Alabama and Georgia, northern part of Arizona and the southern part of Louisiana. In addition, there are very low case rates in counties within Montana, Maine and Utah. *Figure 2* below represents the COVID-19 related death rates for United States counties. There are high death rates in New Jersey counties, counties on the east side of Pennsylvania, southern New York counties and in Massachusetts counties. Furthermore, one can notice very low death rates in counties within Montana, Wyoming and Utah. Counties that have high case and/or death rates could be due to multiple factors that will be explored through this project.

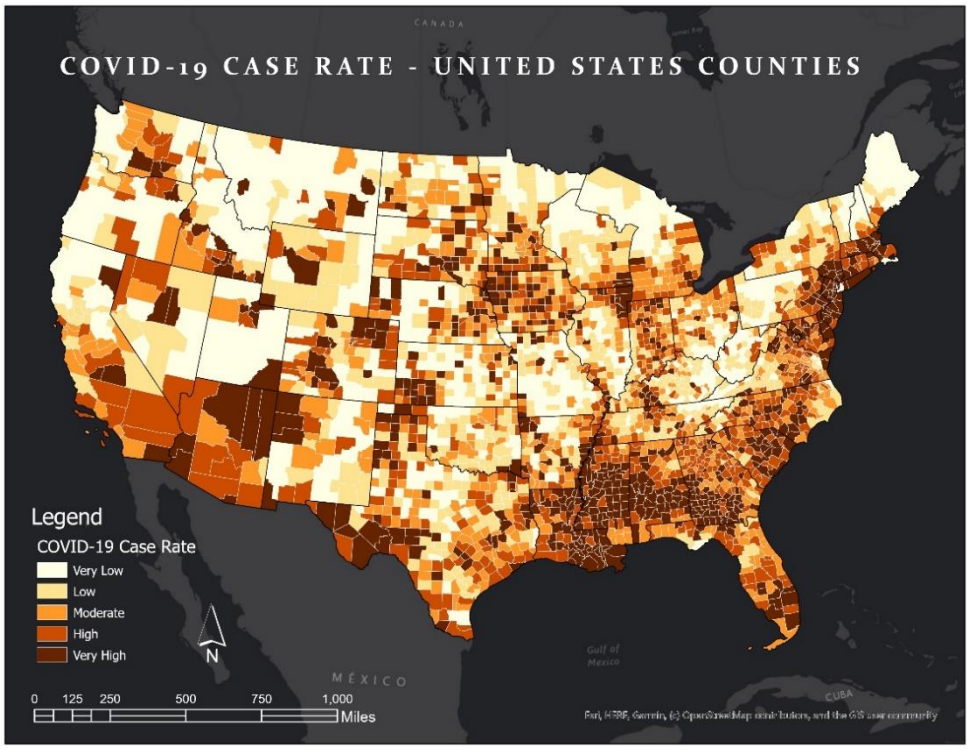


***Figure 1**: COVID-19 Case Rate in the United States counties (48 mainland counties). Case rate is calculated as total number of cases in the counties divided by the population of the county and then it is divided by the number of days since the first COVID-19 case in the county.*

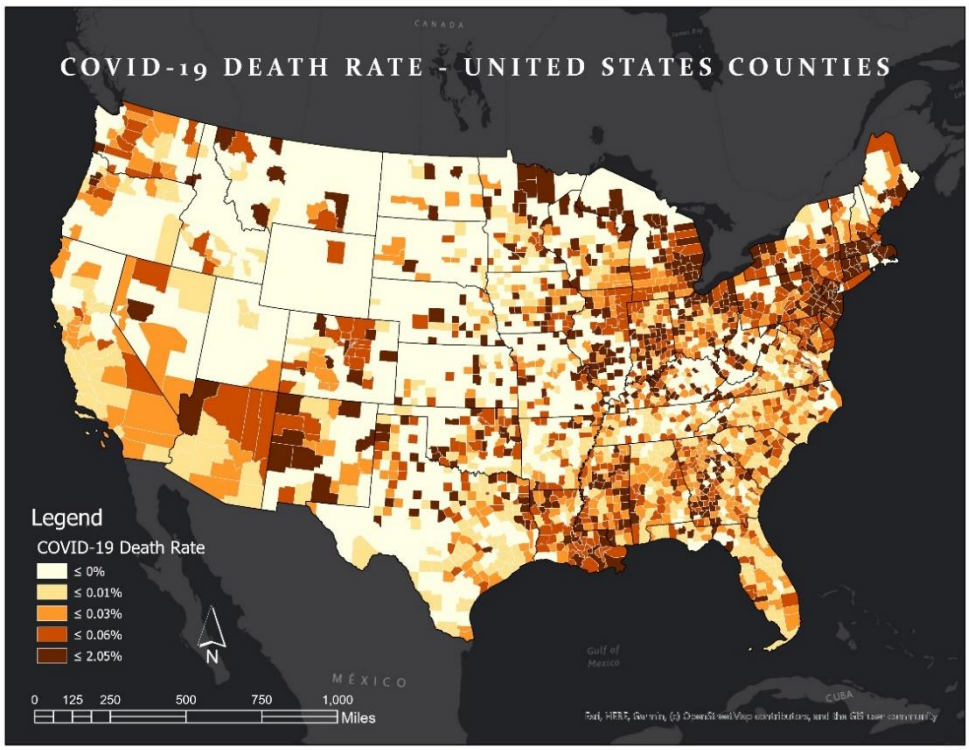


***Figure 2**: COVID-19 Death Rate in the United States counties (48 mainland counties). Death rate is calculated as total number of COVID-19 related death in the county divided by the COVID-19 cases in the county and then it is divided by the number of days since the first COVID-19 case in the county.*

When creating a Vulnerability Index, all the factors and/or criteria that are utilized for it need to be in a rate or percentage format rather than counts or values. Therefore, before beginning the analysis process and examining the various factors, the datasets that were compiled needed to be formatted correctly. Excel and R were utilized to compile and format the datasets to the required layout. The COVID-19 dataset obtained from GitHub included count values for COVID-19 cases and deaths on a county by county bases and those needed to be converted to rates. In addition, when creating index and performing such analysis it is essential to utilize data collected over a long period of time (such as a year), however, since COVID-19 datasets are relatively new, the data is only available from January 2020 to the present date. Thus, for the rate calculation to be accurate, the number of days since each county first had a COVID-19 case needed to be measured. To calculate the case rate, the total number of cases in each county was divided by the population of the county (2019 population attained from American Community Survey) and then the value was divided by the number of days since the first COVID-19 case within that particular county. The death rate was calculated in a similar format, where the total number of COVID-19 deaths in each county was divided by the total number of cases in the county and then that value was divided from the number of days since the first COVID-19 confirmed case within the county. Population density per square mile, percent population over the age of 65, percent population who commute to work and percent population below the federal poverty level were already in a rate format and thus did not need to be modified further. Median household income was a value for each county and thus to format it into a rate for the Vulnerability Index creation, it needed to be standardized. The Median household income was formatted into a rate using a standardization scaling technique that utilized the mean and standard deviation of all the values for the formula.

Furthermore, all health conditions data was based on percent values and thus it was already in a rate format. *Table 1* showcases the variables that are to be included in creating the COVID-19 Vulnerability Index. When the variables were not in a rate or percentage measure they were converted to the correct format using standardization and normalization. In addition when the variable's values are in the inverse direction of the vulnerability measure, their values were inversed so that it is in the right direction. For example, median household income was converted to a rate format from US dollars using standardization and then the inverse of each value was taken so that high value means more vulnerability and low values means less vulnerability.

Variables For Index	Units	Same/Different direction As Vulnerability
Population Density	Square Mile	Same Direction as Vulnerability
Percent Population Over the age of 65	% Percentage	Same Direction as Vulnerability
Percent Population Below Federal Poverty Level	% Percentage	Same Direction as Vulnerability
Percent Population who take Public Transit to work	% Percentage	Same Direction as Vulnerability
Median Household Income	US Dollars	Inverse Direction from Vulnerability
Percent Population with Poor Health	% Percentage	Same Direction as Vulnerability
Heart Failure (population over 65)	% Percentage	Same Direction as Vulnerability
Chronic Kidney Disease (population over 65)	% Percentage	Same Direction as Vulnerability
Chronic Obstructive Pulmonary Disease (population over 65)	% Percentage	Same Direction as Vulnerability
Coronary Artery Disease (population over 65)	% Percentage	Same Direction as Vulnerability
Prevalence of Diabetes in Adults	% Percentage	Same Direction as Vulnerability
Prevalence of Obesity in Adults	% Percentage	Same Direction as Vulnerability

Table 1: *Showcases the variables that will be utilized in the creating the COVID-19 Vulnerability Index.*

After all the variables or factors are formatted and corrected, they were combined into one dataset using a common field of the County FIPs (Federal Information Processing Standards) code. Then JMP (statistical software) was utilized to analyze the distribution and correlation of the factors and to create the Vulnerability Index values. *Figure 3* below showcases the distributions of each of the variables. Factors such as percent population over the age of 65, percent population below the federal poverty level and percent population who carpool or take public transit have a more positively skewed distribution. This indicates that most of the values are towards the lower end of the scale and there are less high values. For example, this means that there is a lower percentage of people who carpool or take public transit to work across all counties than people who drive to work on their own. Median household income, and all the health factors such as heart failure, chronic kidney disease, COPD, coronary artery disease, diabetes and obesity have relatively normal distributions. Percent poor health represents a relatively normal distribution with a large number of counties that are around the mean value (approximately 18%) with regards to percentage of the population that has poor health. Population density is also heavily positively skewed in its

distribution. This means that across the United States there are a lot more counties with low population density than high population densities.

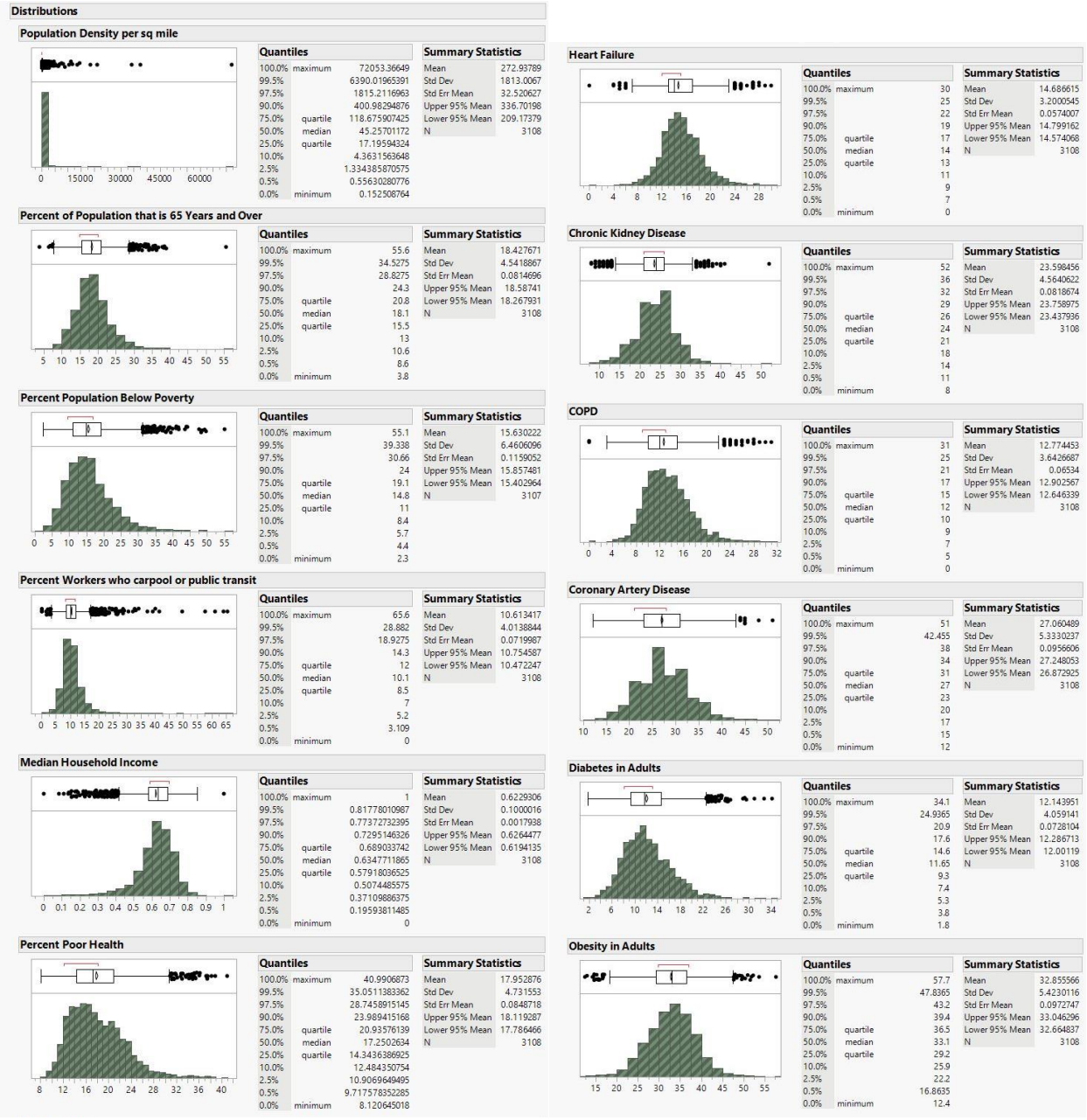


Figure 3*: Distribution of the factors/variables being utilized for the COVID-19 Vulnerability Index.*

The next process in the index creation involved analyzing the correlations in the dataset using multivariate methods to determine the strength of the linear relationship between pairs of variables and the degree to which a pair of variables change together.

Correlations	Population Density per sq mile	Percent of Population that is 65 Years and Over	Percent Population Below Poverty	Percent Workers who carpool or public transit	Median Household Income	Percent Poor Health	Heart Failure	Chronic Kidney Disease	COPD	Coronary Artery Disease	Diabetes in Adults	Obesity in Adults
Population Density per sq mile	1	-0.1231	-0.0088	0.5575	-0.1588	-0.0277	-0.0151	0.0339	-0.0865	0.0006	-0.0848	-0.1531
Percent of Population that is 65 Years and Over	-0.1231	1	-0.1066	-0.1594	0.2727	-0.1534	-0.0801	-0.2645	0.0093	-0.041	0.1133	-0.082
Percent Population Below Poverty	-0.0088	-0.1066	1	0.0924	0.7434	0.819	0.3041	0.3501	0.364	0.1582	0.4035	0.3558
Percent Workers who carpool or public transit	0.5575	-0.1594	0.0924	1	-0.0856	0.0946	-0.0117	-0.0084	-0.0717	-0.0196	-0.0529	-0.1362
Median Household Income	-0.1588	0.2727	0.7434	-0.0856	1	0.6733	0.3335	0.2322	0.4407	0.1414	0.4534	0.4115
Percent Poor Health	-0.0277	-0.1534	0.819	0.0946	0.6733	1	0.4402	0.4771	0.4433	0.3132	0.4702	0.4187
Heart Failure	-0.0151	-0.0801	0.3041	-0.0117	0.3335	0.4402	1	0.5079	0.5148	0.3399	0.3041	0.3226
Chronic Kidney Disease	0.0339	-0.2645	0.3501	-0.0084	0.2322	0.4771	0.5079	1	0.4751	0.2888	0.3642	0.3684
COPD	-0.0865	0.0093	0.364	-0.0717	0.4407	0.4433	0.5148	0.4751	1	0.3286	0.4195	0.3674
Coronary Artery Disease	0.0006	-0.041	0.1582	-0.0196	0.1414	0.3132	0.3399	0.2888	0.3286	1	0.2157	0.1783
Diabetes in Adults	-0.0848	0.1133	0.4035	-0.0529	0.4534	0.4702	0.3041	0.3642	0.4195	0.2157	1	0.5095
Obesity in Adults	-0.1531	-0.082	0.3558	-0.1362	0.4115	0.4187	0.3226	0.3684	0.3674	0.1783	0.5095	1

Table 2*: Pairwise Correlations of the Variables/Factors being utilized for the COVID-19 Vulnerability Index.*

Based on the pairwise correlations results from *Table 2*, there is a strong correlation between percent population who are living in poverty and median household income. This indicates that counties with high median household income tend to have low percentage of people living in poverty and vice versa. In addition there is a relatively strong correlation between population density and percentage of people who carpool or take public transit to work. This indicates that counties with a high population density tend to be the counties where people are more inclined to take public transit or carpool. Moreover, there is a really strong correlation (0.819) between the percent population living in poverty and percent population who have poor health. This indicates that in the counties with high poverty there will also be a high population of people in poor health and in counties with low poverty only minimal percentage of the population will be in poor health. In addition, there is a certain level of correlation between percentage of people living in poverty and all the health factors. Examining the values, it seems that counties with higher level of poverty experience higher health issues among its 65 and older population. Furthermore, there is a positive correlation between each of the health related factors, counties that have a higher prevalence of one health condition tend to have a higher prevalence of the other health conditions and vice versa.

Based on the results from clustered correlations graphic below (*Figure 4*) there is a strong positive correlation between population density and the percent of people who carpool or take public transit to work. This would indicate that as population density increases in a county the percentage of people who take public transit to work increases. Additionally, there is also strong positive correlations between the percentages of the population who are living in poverty, median household income, and the percentage of the population who have poor health. Overall, one can notice a general positive correlation (relationship) between the all the health indicators, median household income, percentage of population with poor health and percent of the population living below poverty level. This indicates that overall in counties across the United States, one will notice

that counties with high poverty rates, low median income and high percentage of people with poor health correlate with high health issues such as the ones considered for the index.

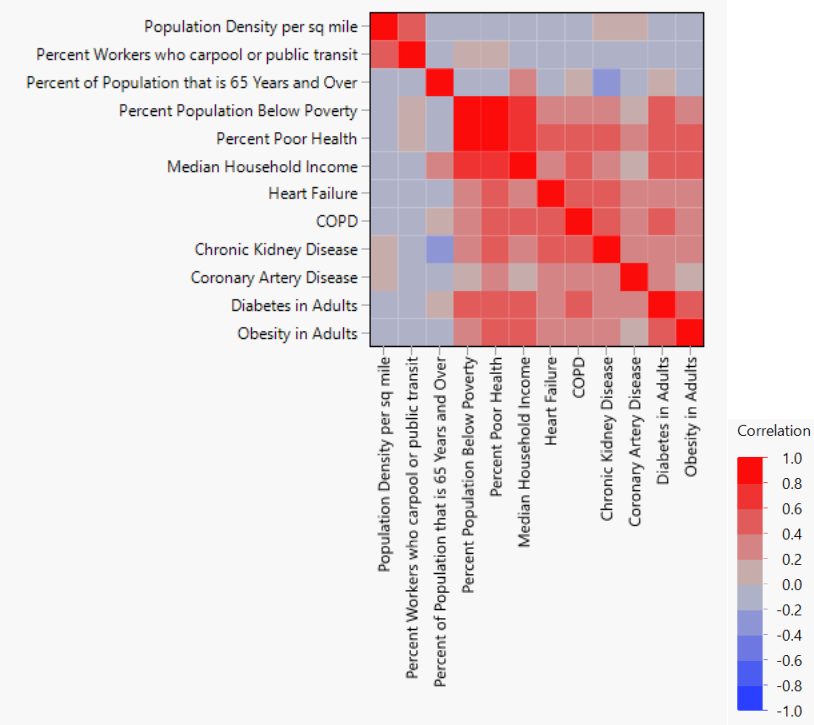


Figure 4*: Clustering the Correlations. 1.0 (Red) indicates a complete positive correlation and -1.0 (Blue) indicates a complete negative correlation.*

Eigenvalues

Number	Eigenvalue	Percent	Cum Percent
1	4.2665	35.554	35.554
2	1.7515	14.596	50.149
3	1.3246	11.038	61.188
4	1.0212	8.510	69.698
5	0.8199	6.832	76.531
6	0.7046	5.872	82.402
7	0.4993	4.161	86.564
8	0.4735	3.946	90.509
9	0.4437	3.697	94.206
10	0.3788	3.157	97.363
11	0.1789	1.491	98.854
12	0.1375	1.146	100.000

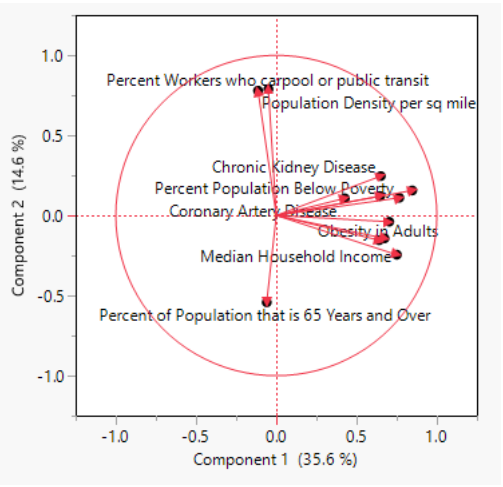


Figure 5*: Eigenvalue Chart and principal components for the 12 variables/factors considered in the COVID-19 Vulnerability Index.*

The next step in the index creation process is to run the principal component analysis on the dataset. Principal component analysis was run in order to create the index values because of the large dataset. Utilizing this method makes the dataset easier to interpret and reducing information loss by maximizing variance (Jolliffe & Cadima, 2016). Based on the Eigenvalue chart in *Figure 5*, the

percentage of variance for the first principal component is 35.6%. In addition, most of the variables in the plot are loading to the right of the axes for principal component 1. This includes all variables other than percent population over the age of 65, percent workers who carpool or take public transit, and population density per square mile. This means that these three variables or factors have the lowest correlations with the other variables considered in the index. This indicates that counties with high population density, high percent population over the age of 65 or high percent people who take public transit will not mean that the counties will have high poverty, many health issues, and low median income. This means that most of the variables have a positive correlation with principal component 1. *Table 3* indicates the weight that each variable will have in calculating the principal component index values. Principal component 1 has the highest weights for the variables. Based on the table, when utilizing principal component 1 for creating the index, percent population with poor health will have the highest weight in the calculation. *Table 4* represents the eigenvector values for principal component 1. These coefficients are utilized to form a linear combination of the original variables and produce principal component variables. The principal component 1 values are utilized to calculate the index value for each county. The index value is calculated as the sum of the each eigenvector value multiplied by the z-score for each variable. The z-score is the raw score of that particular variable minus the mean and then divided by the standard deviation.

Loading Matrix	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12
Population Density per sq mile	-0.11419	0.7791	0.21854	0.29201	0.19823	-0.01859	-0.0236	0.19736	0.34548	-0.21881	0.01432	-0.01844
Percent of Population that is 65 Years and Over	-0.05998	-0.54188	0.43208	0.65007	0.0664	-0.09711	-0.01983	0.15882	0.00565	0.18853	0.11983	0.06398
Percent Population Below Poverty	0.76579	0.11068	0.4278	-0.27786	-0.23866	0.01922	-0.04709	0.02974	0.04868	-0.04256	-0.02273	0.27563
Percent Workers who carpool or public transit	-0.04961	0.78842	0.34603	0.16297	0.09959	0.02743	0.11614	-0.22885	-0.28406	0.26637	-0.0382	0.00074
Median Household Income	0.75037	-0.24482	0.48769	0.01639	-0.14281	-0.07637	0.07801	0.02759	0.10033	0.02138	-0.25552	-0.17322
Percent Poor Health	0.84654	0.1563	0.2232	-0.19205	-0.22012	0.07498	-0.0384	0.05056	-0.05708	-0.0058	0.2932	-0.15776
Heart Failure	0.64651	0.12165	-0.34048	0.23734	-0.08774	-0.3699	0.34741	0.18081	-0.25141	-0.1811	-0.00792	0.03143
Chronic Kidney Disease	0.65067	0.24614	-0.41354	-0.06659	0.10605	-0.20812	-0.33026	0.26118	0.0389	0.323	-0.05381	-0.00136
COPD	0.70331	-0.03875	-0.21648	0.26473	0.01462	-0.27552	-0.09396	-0.50092	0.21848	-0.03104	0.04575	0.01541
Coronary Artery Disease	0.42673	0.10709	-0.38297	0.43767	-0.37309	0.56821	0.01851	0.01432	0.03684	0.03566	-0.05247	0.01469
Diabetes in Adults	0.67338	-0.14275	0.0879	0.12043	0.48191	0.22323	-0.28472	-0.02219	-0.2819	-0.24325	-0.04095	-0.00017
Obesity in Adults	0.64008	-0.15452	-0.10742	-0.18857	0.49987	0.23135	0.39385	0.01364	0.18611	0.15362	0.04028	0.02779

Table 3*: Loading Matrix for the Variables assigns the weights that each variable will have in calculating the matrix.*

Eigenvectors	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12
Population Density per sq mile	-0.05528	0.5887	0.18989	0.28896	0.21893	-0.02215	-0.03339	0.28683	0.51866	-0.35552	0.03386	-0.04971
Percent of Population that is 65 Years and Over	-0.02904	-0.40945	0.37543	0.64327	0.07333	-0.11569	-0.02807	0.23082	0.00848	0.30632	0.28331	0.17254
Percent Population Below Poverty	0.37075	0.08363	0.3717	-0.27495	-0.26357	0.0229	-0.06663	0.04322	0.07309	-0.06915	-0.05375	0.74329
Percent Workers who carpool or public transit	-0.02402	0.59574	0.30066	0.16127	0.10999	0.03268	0.16436	-0.33259	-0.42645	0.43278	-0.09031	0.00199
Median Household Income	0.36328	-0.18499	0.42374	0.01622	-0.15771	-0.09099	0.1104	0.0401	0.15062	0.03473	-0.60411	-0.46711
Percent Poor Health	0.40984	0.1181	0.19393	-0.19004	-0.2431	0.08932	-0.05434	0.07347	-0.08569	-0.00942	0.6932	-0.42541
Heart Failure	0.313	0.09192	-0.29584	0.23486	-0.0969	-0.44066	0.49164	0.26277	-0.37744	-0.29424	-0.01872	0.08475
Chronic Kidney Disease	0.31501	0.18599	-0.35932	-0.0659	0.11712	-0.24794	-0.46737	0.37958	0.05839	0.5248	-0.12721	-0.00366
COPD	0.3405	-0.02928	-0.1881	0.26196	0.01615	-0.32823	-0.13296	-0.72799	0.328	-0.05043	0.10817	0.04155
Coronary Artery Disease	0.2066	0.08092	-0.33275	0.4331	-0.41203	0.67691	0.02619	0.02082	0.05531	0.05795	-0.12404	0.03962
Diabetes in Adults	0.32601	-0.10787	0.07637	0.11918	0.53222	0.26593	-0.40292	-0.03225	-0.42321	-0.39523	-0.09682	-0.00045
Obesity in Adults	0.30988	-0.11676	-0.09334	-0.1866	0.55205	0.27561	0.55736	0.01982	0.27941	0.24959	0.09523	0.07494

Table 4*: Eigenvectors for the Variables/Factors that are utilized in creating the COVID-19 Vulnerability Index.*

Results

The results compare the final vulnerability index values to the COVID-19 case rates and death rates. Spatial clustering was performed utilizing GeoDa on the case rates and death rates datasets using a queen contiguity spatial clustering method. *Figure 6* is the cluster map for case rates in United States counties. High-high clusters mean that there are high case rates in those counties and the counties surrounding (neighboring) it. Counties in the southern part of New York, eastern part of Pennsylvania, in New Jersey and in the central counties of Georgia, Alabama and Mississippi have high case rates surrounded by high case rates. The low-low clusters indicate that those counties with low case rates are surrounded by other counties with low case rates. In Montana there are low case rate counties neighboring other low case rate counties. *Figure 7* represents the significance map for case rates which showcases the deviation of county case rates from a random pattern. A p value of 0.001 is the most significant, showing the locations with a significant local spatial auto-correlation (local Moran's I). As the shades of green get darker the degree of significance increases. Based on the map, it can be noted that counties in the state of Montana have a high significant location spatial auto-correlation, indicating that there is a stronger relationship between those counties and its neighbors.

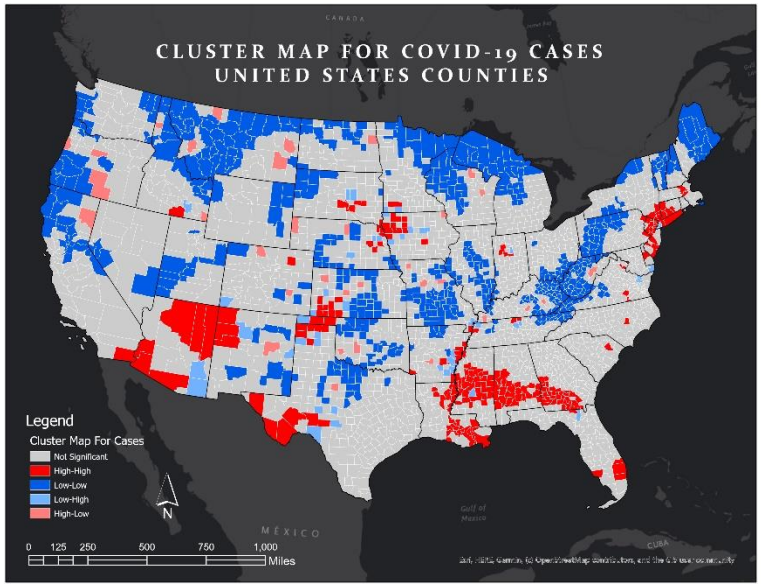


Figure 6: *Cluster map for Case Rates per United States County*

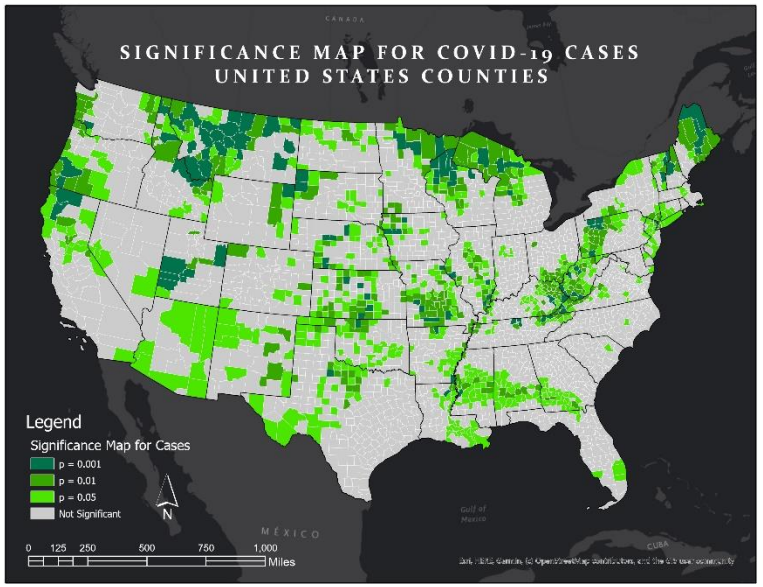


Figure 7: *Significance map for Case Rates per United States County*

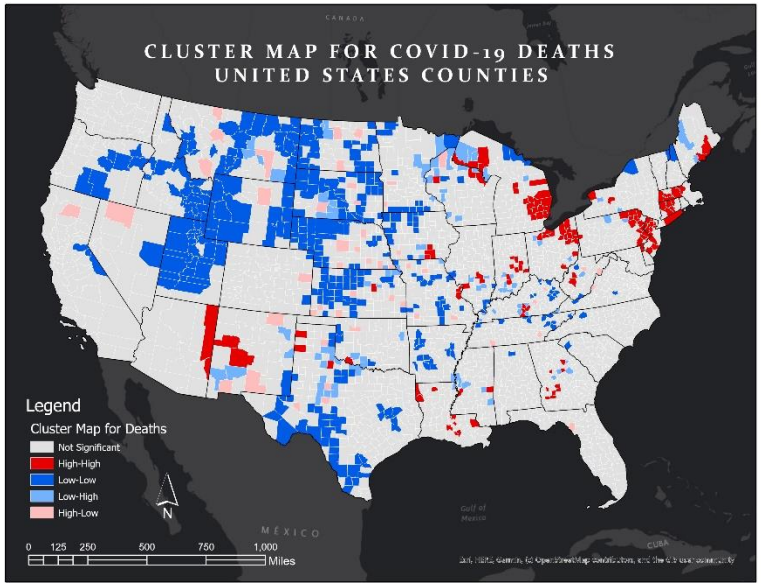


Figure 8: *Cluster map for Death Rates per United States County*

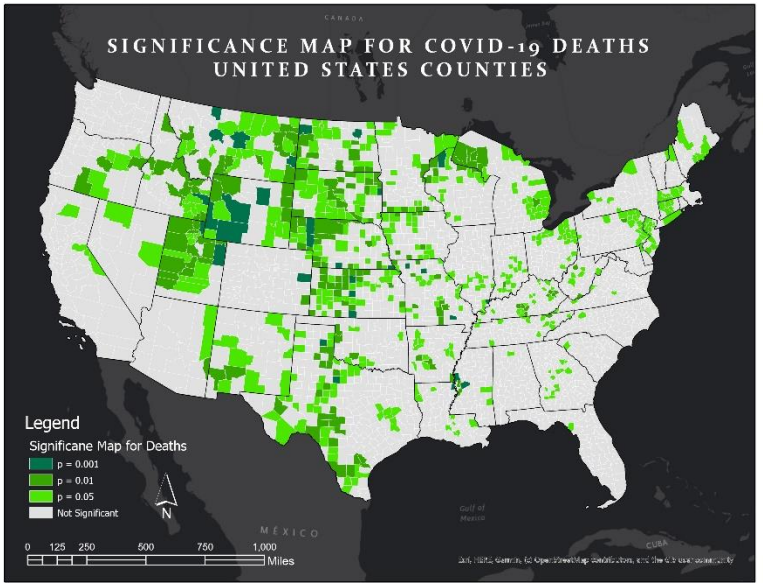


Figure 9: *Significance map for Death Rates per United States County*

Figure 8 is the cluster map for death rates in United States counties. With regards to death rate, there are very few with high death rates that are surrounded by other counties with high death rate. One can notice a few of such counties in Massachusetts and Pennsylvania. On the other hand there are a lot of counties (for example counties in Montana, Wyoming and Utah) with low death rates that are surrounded by other counties with low death rates. Based on the significance map for death rates from *Figure 9*, one can notice that there not a lot of spatial auto-correlation between counties and its neighbors other than in most counties within Montana, Wyoming, Utah and a few others. Comparing the cluster maps for case rates and death rates there are not many similarities in the spatial clustering of case rates and death rates. The only region where there are high-high values for the case rates cluster map and the death rates cluster map is in the New York and New Jersey area.

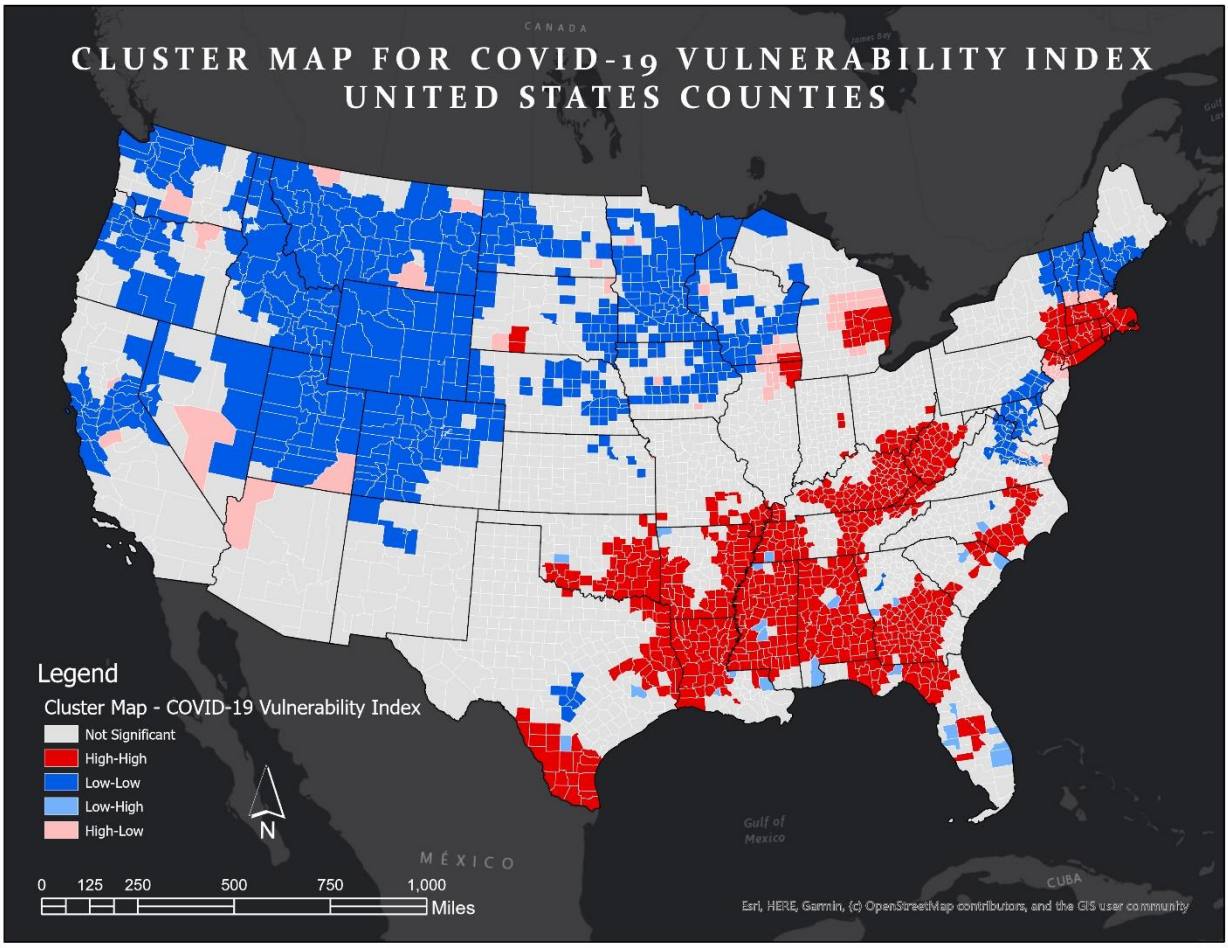


***Figure 10**: Cluster Map representing the COVID-19 Vulnerability Index results.*

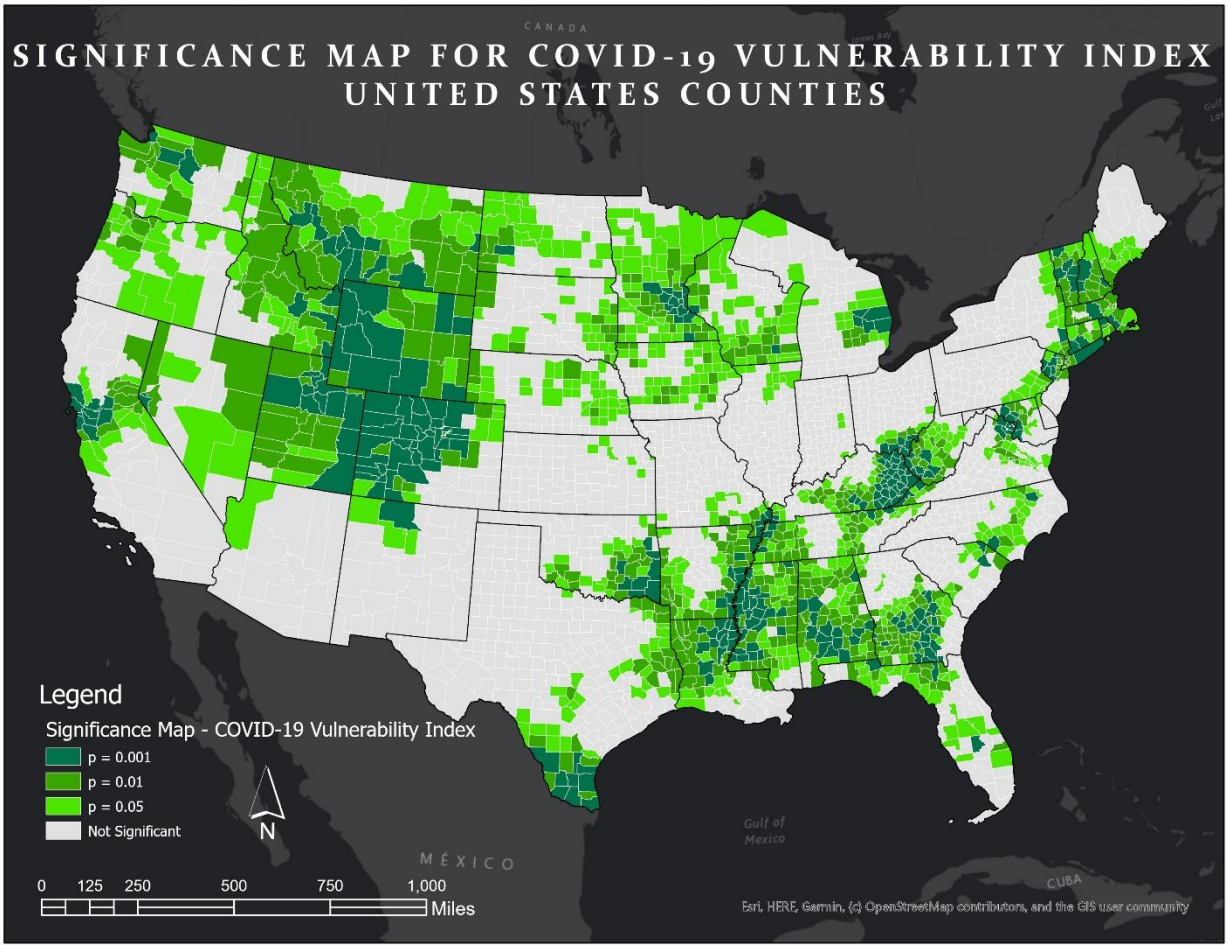


Figure 11*: Significance Map representing the COVID-19 Vulnerability Index results.*

Figure 10 and *Figure 11* represent the Cluster and Significance map results from the Vulnerability Index values. Based on the results south eastern counties in New York, northern counties in New Jersey, and all counties in Connecticut, Rhode Island and Massachusetts are counties that are highly vulnerable and they are surrounded by other highly vulnerable counties. Factors that would lead to the high vulnerability in these counties would be the high population density and large number of people who commute to work. In addition, many counties in the south eastern part of the United States are also identified to be highly vulnerable and surrounded by other highly vulnerable counties. This includes counties in Florida, Georgia, Alabama, Mississippi, Louisiana, Arkansas, Tennessee, Kentucky and West Virginia. These counties have high vulnerability due to factors such as percent population poor health, pre-existing health condition, and high poverty. On the other hand, counties in Minnesota, Montana, Idaho, Wyoming, Utah and Colorado show low vulnerability and have a high spatial auto-correlation (relationship) with neighboring counties.

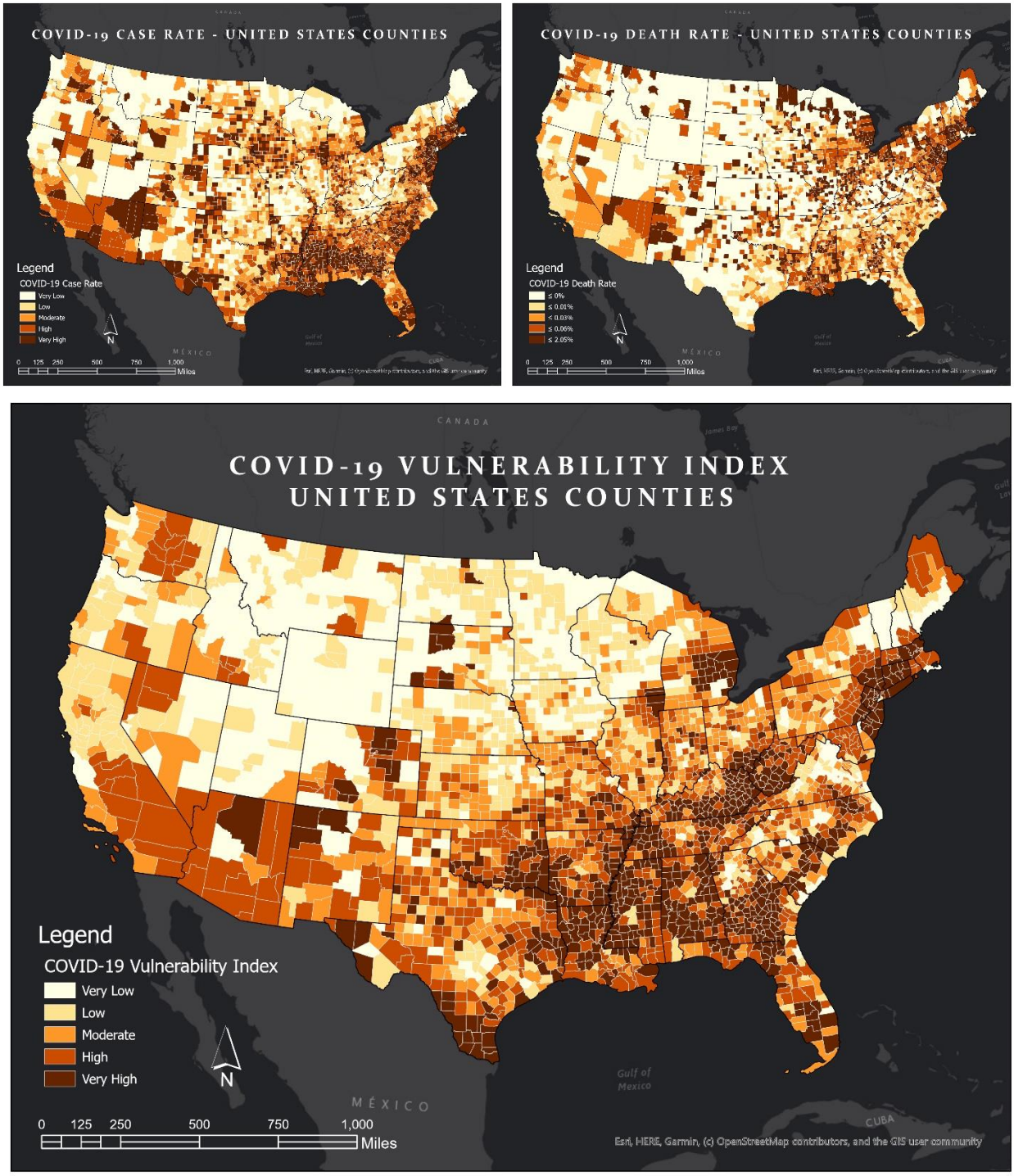


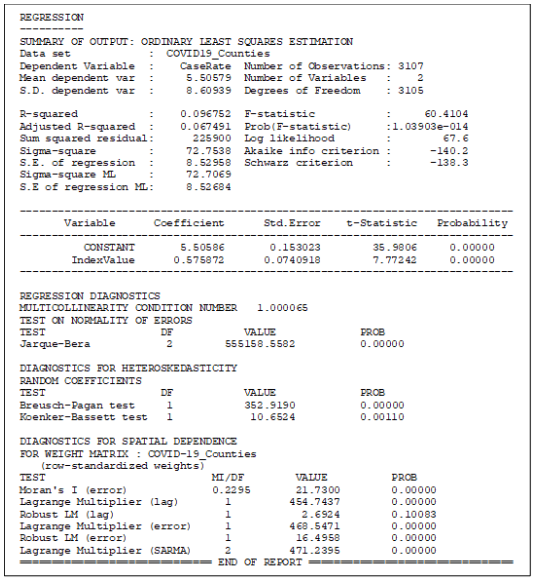
Figure 12: *Comparing the COVID-19 Case Rate, COVID-19 Death Rate and COVID-19 Vulnerability Index results.*

Comparing the maps from *Figure 12*, it can be concluded that the COVID-19 Vulnerability Index results correlate more with the COVID-19 case rates. Counties that have a high Vulnerability Index value tend to have high case rates in actuality. This indicates that the COVID-19 Vulnerability Index is a good method to assess the case rate risk within United States counties. On the other

hand, counties that are identified to be high risk based on the COVID-19 Vulnerability Index do not necessarily correlate with the counties that have high death rates in reality. However, counties that are at low risk based on the COVID-19 Vulnerability Index tend to have low death rates in actuality. Based on the analysis conducted the COVID-19 Vulnerability Index is a good measure for identifying counties that are at risk for increasing number of cases however it is not necessarily a good measure for identifying counties that are at high risk for deaths due to COVID-19.

Discussion

To evaluate how valuable the COVID-19 Vulnerability Index is in identifying counties that are at risk of high COVID-19 case rates, regression models need to be applied. GeoDa was utilized to run the regression models. An ordinary least square estimation regression model was run utilizing the COVID-19 Vulnerability Index values and the COVID-19 case rates (*Figure 13)*. Since the Prob (F-statistic) is less than 0.05 it can be concluded that the model tests are significant. The adjusted r-squared value is 6.7%, meaning that 6.7% of the variation in the Vulnerability Index values can be accounted for by the case rates. This was determined because the adjusted r-squared value identifies the level of variation in the dependent variable (Vulnerability Index value) that can be accounted for by the independent variable (case rate). In addition, the p-value of the Vulnerability Index indicates that its coefficient is statistically important.



```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  COVID19_Counties
Dependent Variable  :    CaseRate  Number of Observations: 3107
Mean dependent var  :     5.50579  Number of Variables   :    2
S.D. dependent var  :     8.60939  Degrees of Freedom    : 3105

R-squared           :    0.096752  F-statistic           :     60.4104
Adjusted R-squared  :    0.067491  Prob(F-statistic)     :1.03903e-014
Sum squared residual:      225900  Log likelihood        :        67.6
Sigma-square        :     72.7538  Akaike info criterion :      -140.2
S.E. of regression  :     8.52968  Schwarz criterion     :      -138.3
Sigma-square ML     :     72.7069
S.E of regression ML:     8.52684

------------------------------------------------------------------------------
      Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
      CONSTANT         5.50586       0.153023        35.9806       0.00000
     IndexValue       0.575872       0.0740918        7.77242      0.00000
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER    1.000065
TEST ON NORMALITY OF ERRORS
TEST                    DF           VALUE              PROB
Jarque-Bera              2       555158.5582           0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                    DF           VALUE              PROB
Breusch-Pagan test       1           352.9190           0.00000
Koenker-Bassett test     1            10.6524           0.00110

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : COVID-19_Counties
   (row-standardized weights)
TEST                           MI/DF        VALUE          PROB
Moran's I (error)              0.2295       21.7300        0.00000
Lagrange Multiplier (lag)           1      454.7437        0.00000
Robust LM (lag)                     1        2.6924        0.10083
Lagrange Multiplier (error)         1      469.5471        0.00000
Robust LM (error)                   1       16.4958        0.00000
Lagrange Multiplier (SARMA)         2      471.2395        0.00000
======================== END OF REPORT ========================
```

Figure 13*: Ordinary Least Square Estimation Regression Model for the COVID-19 Case Rate and Vulnerability Index values.*

From the diagnostics for spatial dependence for weight matrix, one can conclude that the Lagrange Multiplier Test for lag and error are significant. Thus, the spatial lag model for maximum likelihood estimation was also run. The results from the spatial lag model are represented in *Figure 14*. The r-squared value in the spatial lag model is 67.5 %. This means that based on case rate, the spatial lag model explains 67.5% of the variance in the Vulnerability Index values. In addition, the spatial lag value for case rates is statistically significant. Due to the high r-squared value and high log likelihood the spatial lag model is a good regression method in assessing COVID-19 case rates using the COVID-19 Vulnerability Index values within United States counties.

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : COVID-19_Counties
Spatial Weight      : COVID-19_Counties
Dependent Variable  :    CaseRate  Number of Observations: 3107
Mean dependent var  :     5.50579  Number of Variables   :    3
S.D. dependent var  :     8.60939  Degrees of Freedom    : 3104
Lag coeff.   (Rho)  :    0.456586

R-squared           :    0.675361  Log likelihood        :         192
Sq. Correlation     : -           Akaike info criterion :     -267.47
Sigma-square        :     62.3452  Schwarz criterion     :     -255.02
S.E of regression   :      7.8959

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value     Probability
------------------------------------------------------------------------------
      W_CaseRate      0.456586      0.0230242        19.8307     0.00000
        CONSTANT       3.00952       0.190688        15.7824     0.00000
      IndexValue      0.433413      0.0697113        6.21726     0.00000
------------------------------------------------------------------------------


REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                DF      VALUE        PROB
Breusch-Pagan test                  1       319.3112     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : COVID-19_Counties
TEST                                DF      VALUE        PROB
Likelihood Ratio Test               1       351.1999     0.00000
========================= END OF REPORT =============================
```

Figure 14: *Spatial Lag Model – Maximum Likelihood Estimation Model for the COVID-19 Case Rate and Vulnerability Index values.*

From the analysis carried out in the methodology and the results and findings from the regression models, it can be concluded that the COVID-19 Vulnerability Index is a good measure to identify the at risk counties for high COVID-19 cases in the United States. However, the results associating the COVID-19 Vulnerability Index to the death rates were not very conclusive Thus, the Vulnerability Index will not be useful in determining and identifying counties that have a high risk of increasing death rates due to COVID-19. The main reason for this could be the size of the death counts dataset. If more COVID-19 related deaths occur over a longer period of time, one would have a more accurate death rate based dataset for COVID-19, using normalization that is more suitable.

Even though the COVID-19 Vulnerability Index is a relatively good measure in identifying counties at risk of increasing case rates, the accuracy of the index could definitely be improved. Looking back at *Figure 4*, that clustered the correlations between the multiple factors and variables that were being utilized in the creation of the index, one can conclude that there are multiple factors that do not have a strong correlation with each other. There is a strong divide between the health related vulnerability factors and the socio-economic, demographic and transportation related vulnerability factors. One way to improve the index and its accuracy in identifying at risk counties would be to create separate index values for different types of factors and then combine the index values. This would mean that a Vulnerability Index would be created for all the pre-existing health conditions as they are highly correlated with each other and would provide more accurate results that are not tainted by factors that do not have a strong relationship with them. Then a Vulnerability Index will be created for socio-economic factors such as percent population living below poverty, median household income and more. This process can be carried out for all the different type of factors. Then weights can be assigned to the resulting index values based on the type of factor and those index values can then be combined to create the final COVID-19 Vulnerability Index. This is a highly likely scope for additional research into to the creation of COVID-19 Vulnerability Index.

Furthermore, the COVID-19 Vulnerability Index can be improved by including more factors that impact the number of COVID-19 cases and deaths in a county. This would comprise of health infrastructure data that includes factors such as number of hospital beds per county, number of ICU beds per county, percent population that lives in nursing homes and community homes for the elderly and more. It would have been ideal to include these factors in the project however, it was not possible to attain accurate and up to date data for those factors in every county. Moreover, for further research into the creation of the COVID-19 Vulnerability Index it would be great include demographic factors such as percent population of minority ethnicity. Based on documents from the Center for Disease Control, such demographic factors can also contribute to the number of COVID-19 cases rates (Centers for Disease Control and Prevention, 2020).

Reviewing the results, it can be concluded that factors such as population density and percent population who commute to work had the most impact on the number of cases in areas such as

New York, New Jersey, Pennsylvania and Massachusetts. This is because in the counties within those areas there is a large population density. Furthermore, in southern states, health related factors had the most impact on the number of COVID-19 related cases. In counties in Georgia, Alabama, Florida, Tennessee and more, there are a much higher percentage of people with pre-existing health conditions that impact the severity of COVID-19 and thus those factors were detrimental to the increasing case rate and death rates in those counties. Ultimately, looking at the different counties through the 48 states considered in the study, certain counties had higher cases rates due to certain factors and variables and other counties had high case rates due to completely different factors. This also applies to the counties with low COVID-19 case rates as some of those counties and states were places with very low population density and some of the counties were places with a relatively lower percentage of population with poor health and other pre-existing conditions.

Conclusion

Since the beginning of 2020, COVID-19 has become the center of people's lives in terms of communication/news, career, travel, health and more. It is thus, very important to evaluate the risk of places and people to COVID-19. The aim of this project was to develop a COVID-19 Vulnerability Index that can be applied to counties within the United States to identify counties that are more vulnerable to COVID-19 cases and COVID-19 related deaths.

The study began by identifying a set of factors that impact the COVID-19 cases and related deaths through a literature review of previous studies exploring factors that have an impact on COVID-19 spread and vulnerability. From the literature review, 12 factors were identified, obtained, examined and organized for the Vulnerability Index creation process. The 12 factors were then analyzed to determine their distribution and correlation. Utilizing Principal Component Analysis technique with eigenvectors, a Vulnerability Index was created based on factors that were more correlated to each other having more weight in the index value. The case rates, death rates and the COVID-19 Vulnerability Index values were further analyzed using spatial clustering to determine how neighboring counties influence values. Then the case rates, death rates and COVID-19 Vulnerability Index value results were mapped and compared to assess if the Vulnerability Index values correlated with the case rates and death rates. Based on the findings, it was determined that

the COVID-19 case rates correlated well with the Vulnerability Index results indicating that the index values could be utilized to determine counties that are at risk of increases COVID-19 cases. To further evaluate those findings, spatial regression models were run including the ordinary least squares model and the spatial lag model. Based on the results from the spatial lag model it can be concluded that the COVID-19 Vulnerability Index is a good measure of COVID-19 cases within counties in the United States.

Further research into the creation of the COVID-19 Vulnerability Index can potentially include additional factors such as health infrastructure and ethnicity related demographic factors to develop a more accurate COVID-19 Vulnerability Index. Moreover, the factors utilized in the creation of the index can be divided into separate indexes (based on the type of factor) and then combined together to create a more accurate index.

References

Centers for Disease Control and Prevention. (2020, April 29). *Cases in the U.S.* Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html

Centers for Disease Control and Prevention. (2020, June 25). *Coronavirus Disease 2019 (COVID-19): Older Adults*. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/coronavirus/2019-ncov/specific-groups/high-risk-complications/older-adults.html

Chandrasekhar, R., Sloan, C., Mitchel, E., Ndi, D., Alden, N., Thomas, A., . . . Lindegren, M. (2017). Social determinants of influenza hospitalization in the United States. *Influenza Other Respi Viruses*, 479-488.

Epidemiology Working Group for NCIP Epidemic Response. (2020). The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19) in China. *Zhonghua Liu Xing Bing Xue Za Zhi*, 145-151.

Henry, B. M., & Lippi, G. (2020). Chronic kidney disease is associated with severe coronavirus disease 2019 (COVID-19) infection. *International Urology and Nephrology*, 1193-1194.

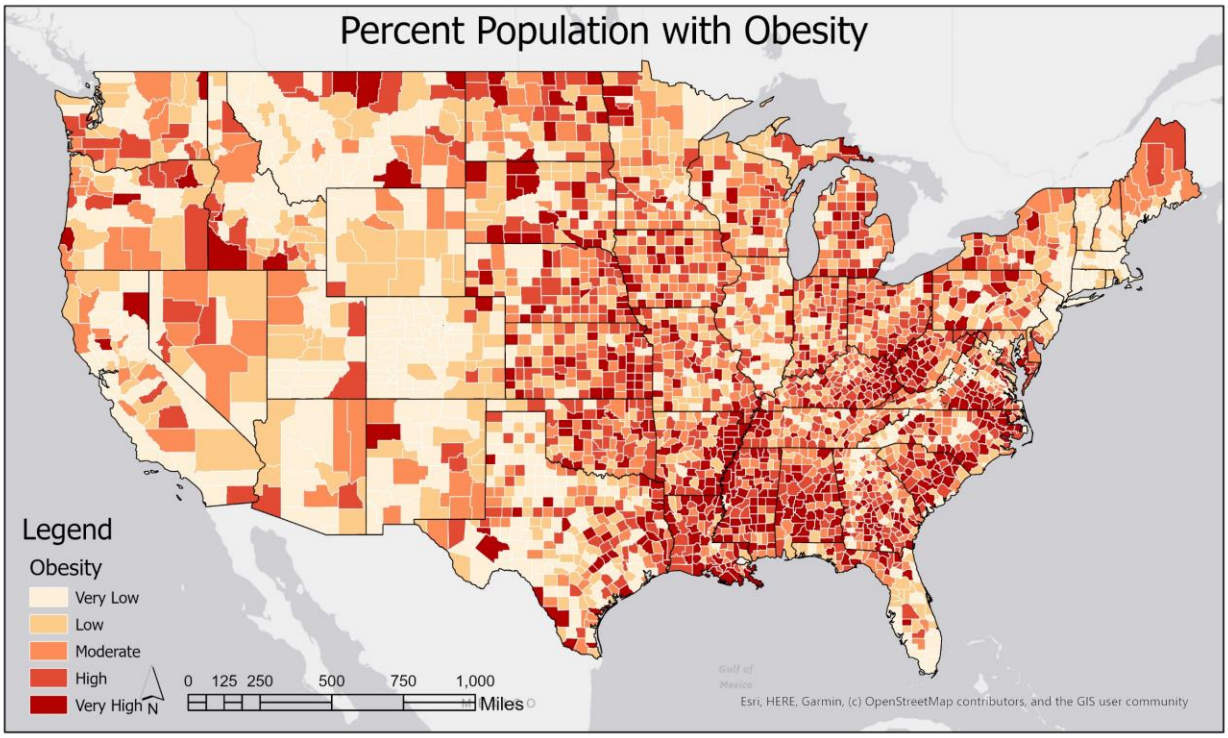
Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 1-16.

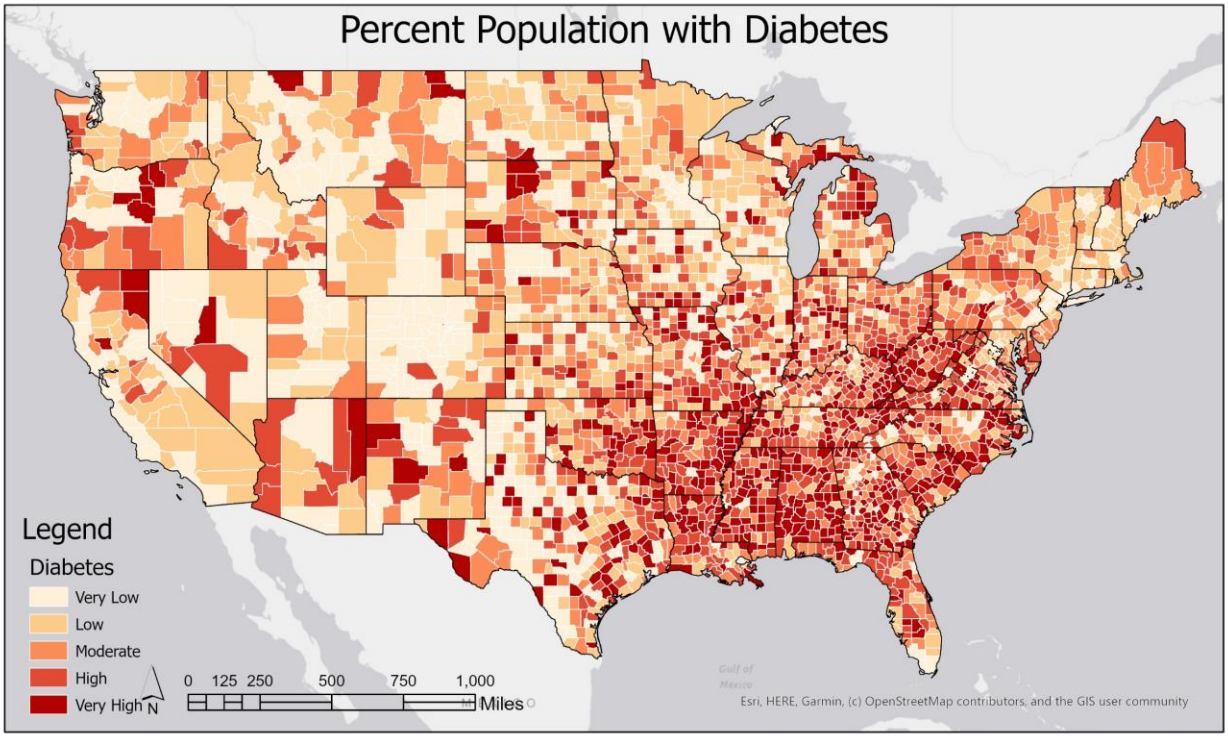
Ramírez, I. J., & Lee, J. (2020). COVID-19 Emergence and Social and Health Determinants in Colorado: A Rapid Spatial Analysis. *Internation Journal of Environmental Research and Public Health*, 1-15.

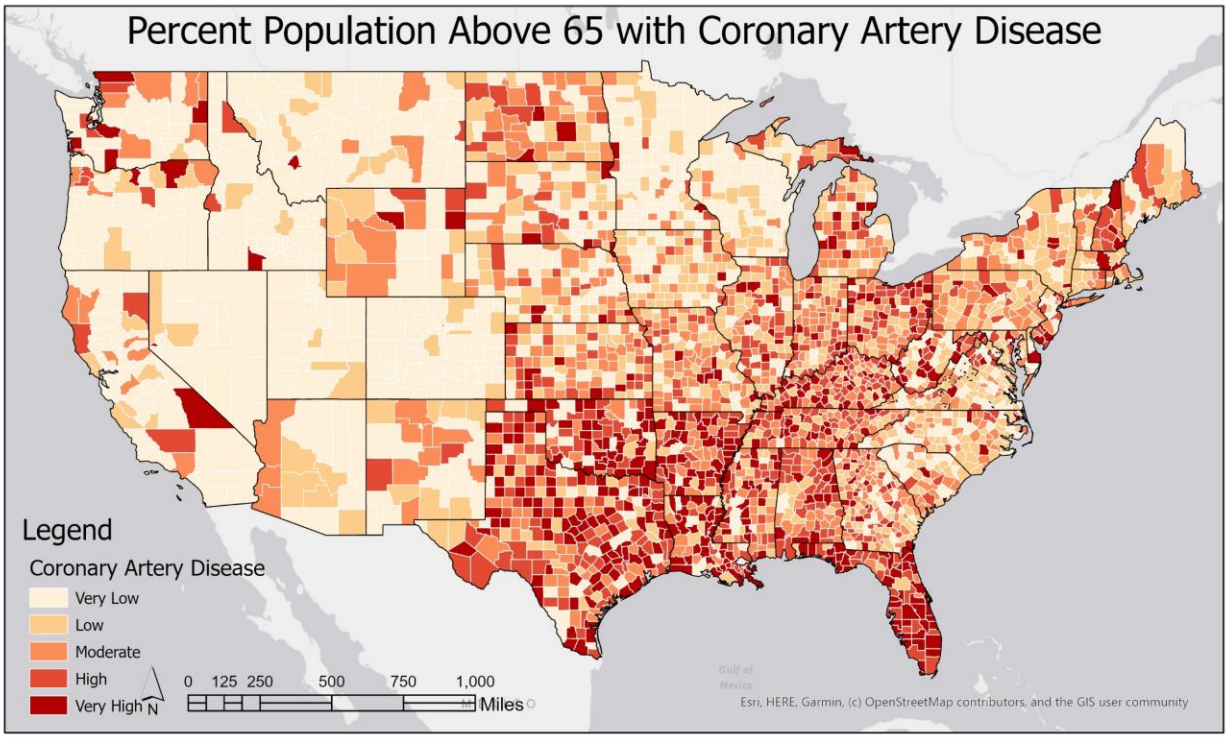
Savini, L., Candeloro, L., Calistri, P., & Conte, A. (2020). A Municipality-Base Approach Using Commuting Census Data to Characterize the Vulnerability to Influenza-like Epidemic: The COVID-19 Application in Italy. *Microorganisms*, 1-20.

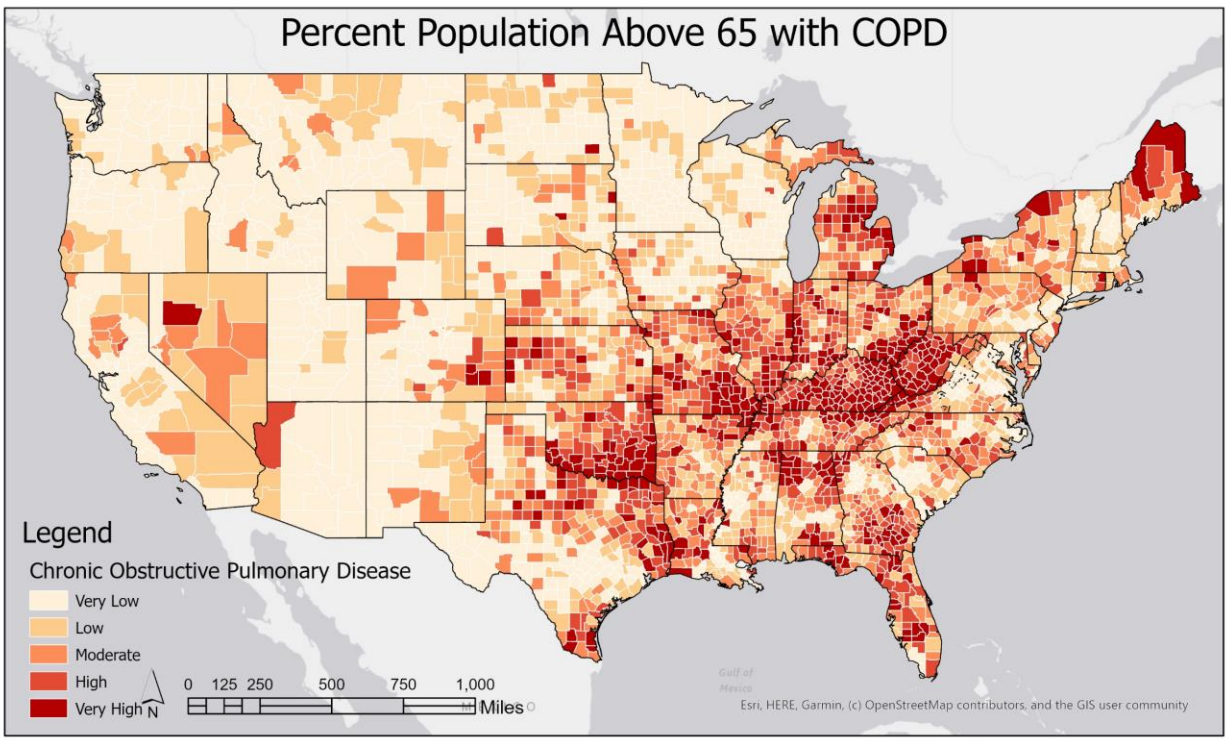
Appendix

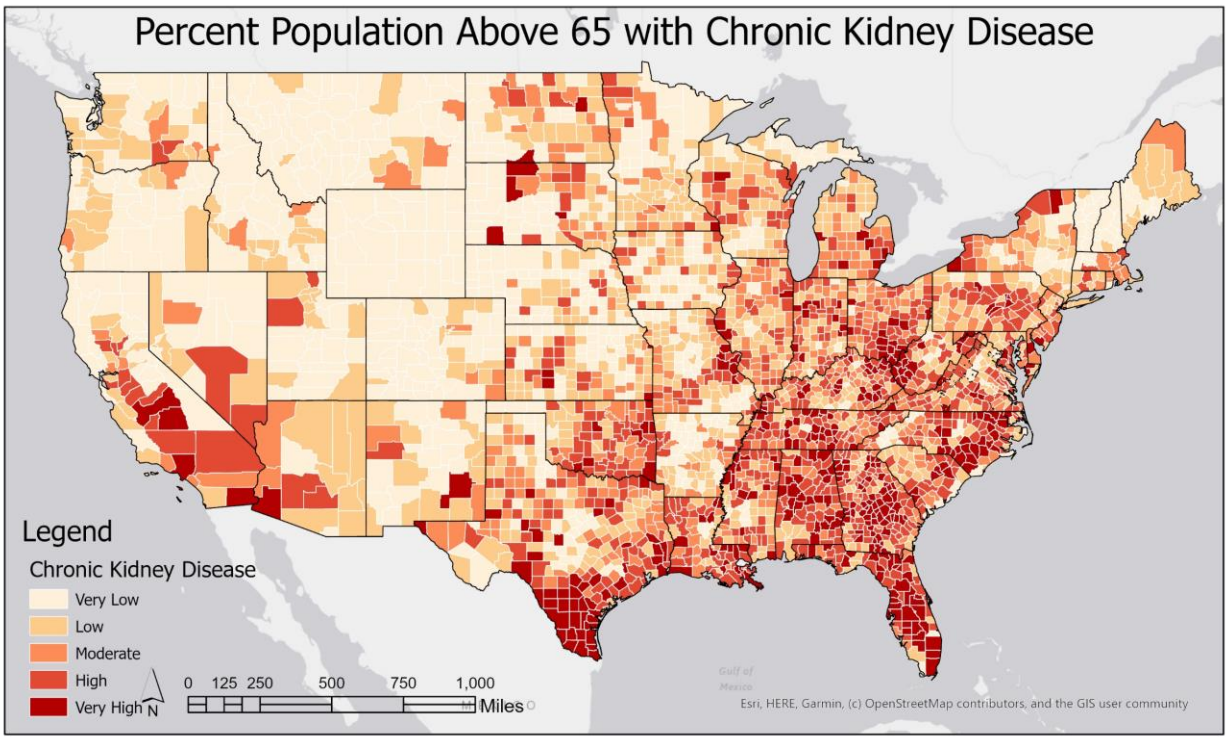
Additional Maps

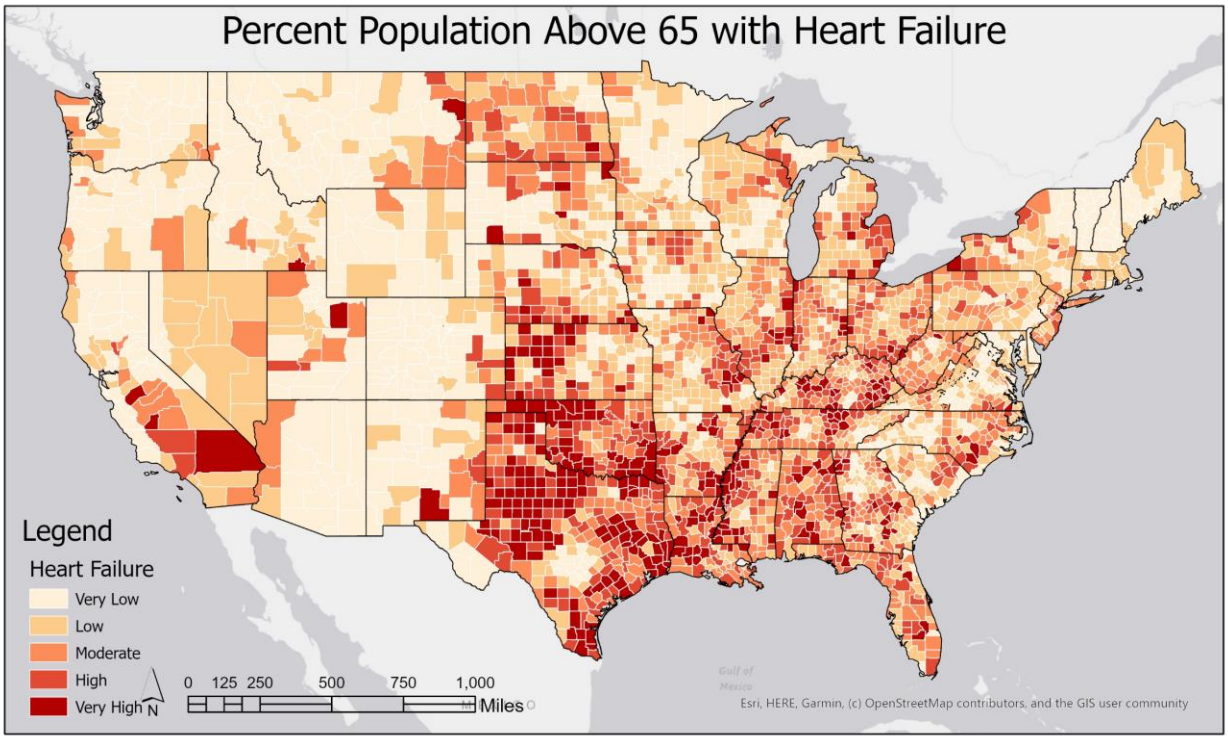


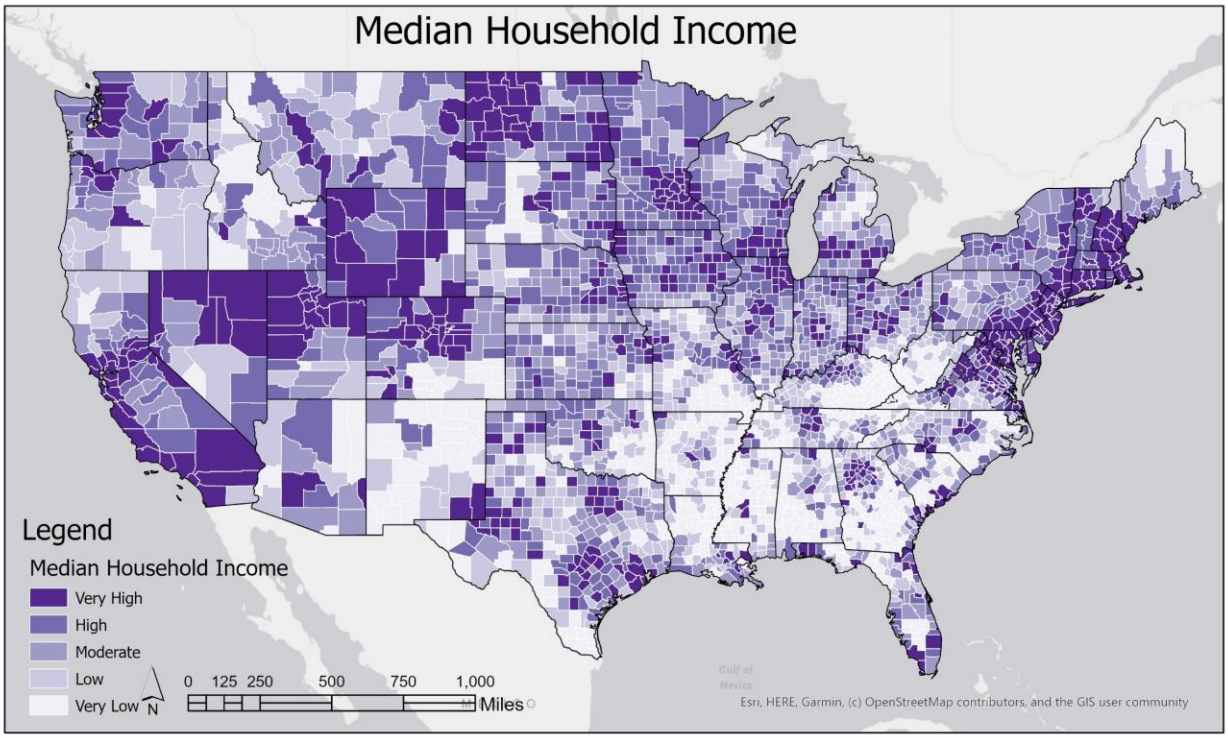


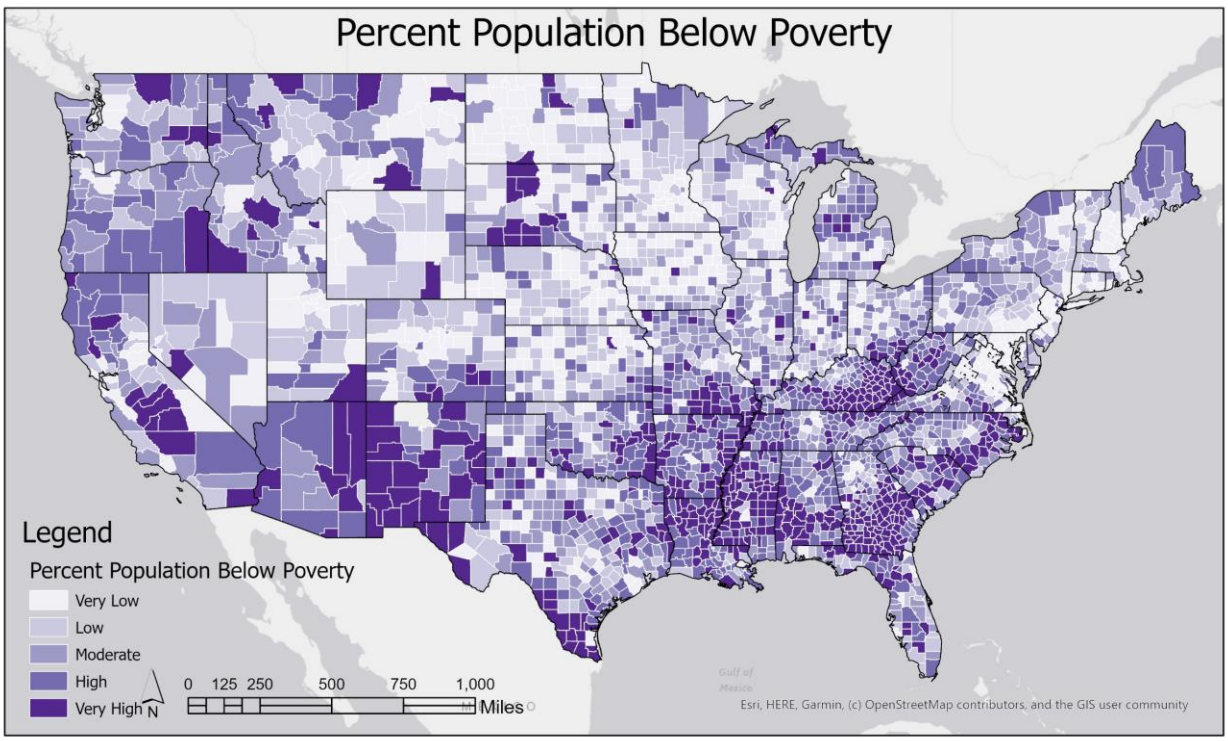

Percent Population Above 65 with Coronary Artery Disease
Legend
Coronary Artery Disease
Very Low
Low
Moderate
High
Very High
N
0 125 250 500 750 1,000
Miles
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

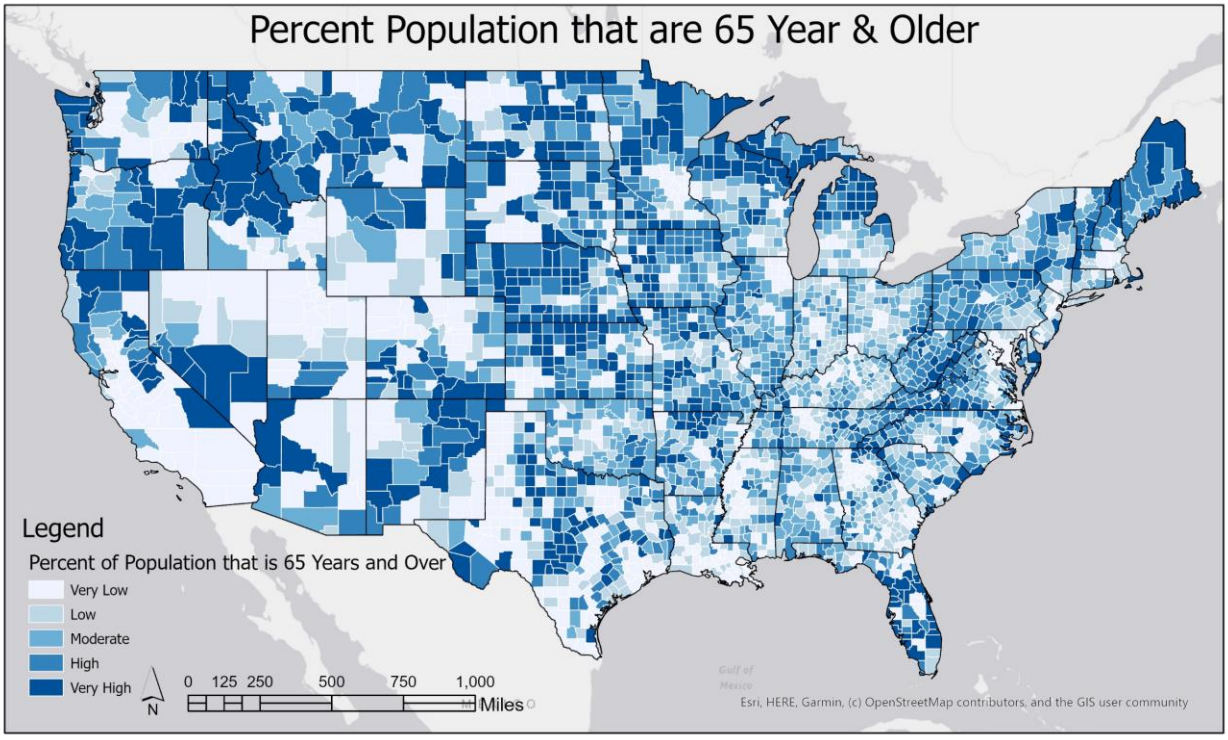

Percent Population Above 65 with COPD
Legend
Chronic Obstructive Pulmonary Disease
Very Low
Low
Moderate
High
Very High
N
0 125 250 500 750 1,000
Miles
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

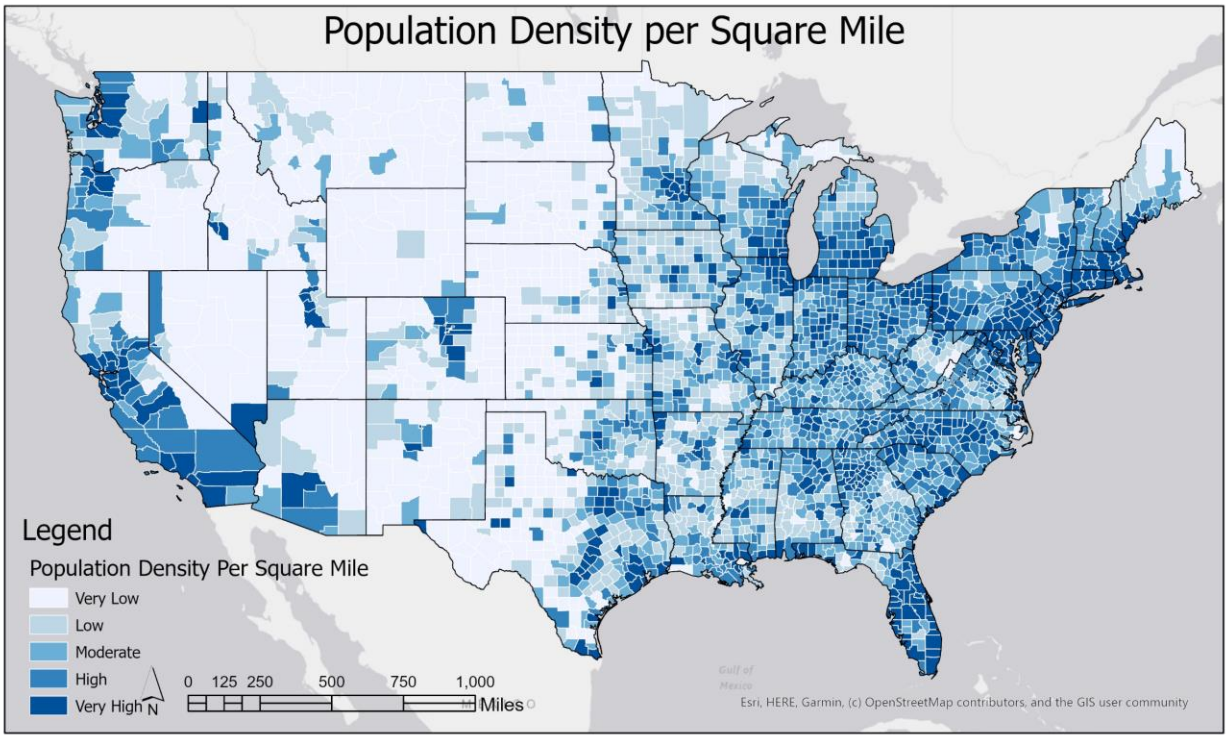

Percent Population Above 65 with Chronic Kidney Disease
Legend
Chronic Kidney Disease
Very Low
Low
Moderate
High
Very High
N
0 125 250 500 750 1,000
Miles
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

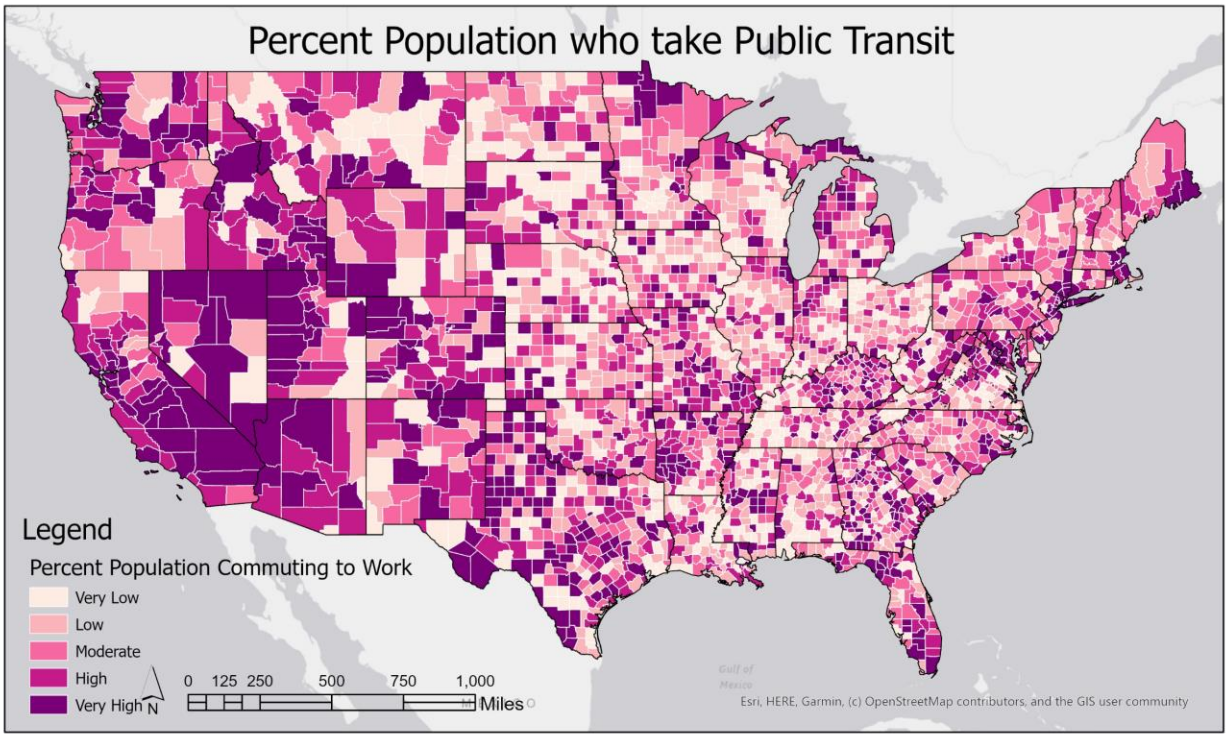

Percent Population Above 65 with Heart Failure
Legend
Heart Failure
Very Low
Low
Moderate
High
Very High
N
0 125 250 500 750 1,000
Miles
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

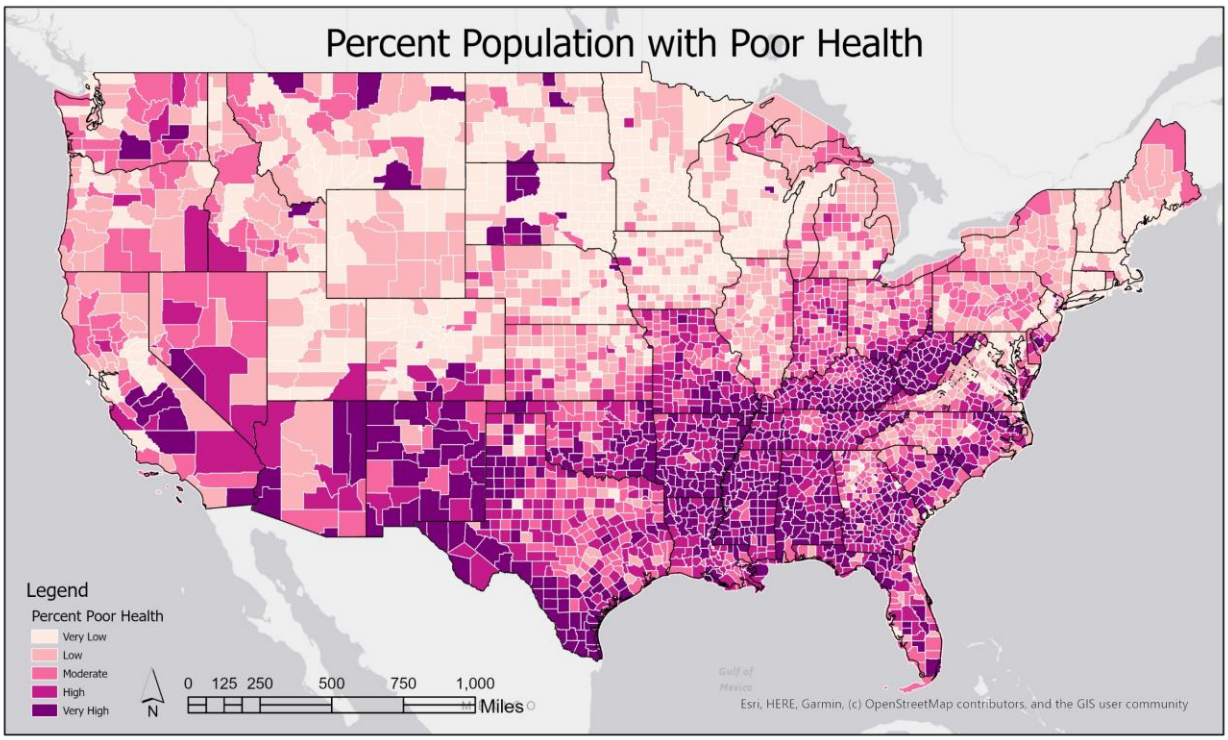

Median Household Income
Legend
Median Household Income
Very High
High
Moderate
Low
Very Low
N
0 125 250 500 750 1,000
Miles
Gulf of Mexico
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community


Percent Population Below Poverty
Legend
Percent Population Below Poverty
Very Low
Low
Moderate
High
Very High
N
0 125 250 500 750 1,000
Miles
Gulf of Mexico
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community


Percent Population that are 65 Year & Older
Legend
Percent of Population that is 65 Years and Over
Very Low
Low
Moderate
High
Very High
N
0 125 250 500 750 1,000
Miles
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community


Population Density per Square Mile
Legend
Population Density Per Square Mile
Very Low
Low
Moderate
High
Very High
N
0 125 250 500 750 1,000
Miles
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community


Percent Population who take Public Transit
Legend
Percent Population Commuting to Work
Very Low
Low
Moderate
High
Very High
N
0 125 250 500 750 1,000
Miles
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community


Percent Population with Poor Health
Legend
Percent Poor Health
Very Low
Low
Moderate
High
Very High
N
0 125 250 500 750 1,000
Miles
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community